Sentiment Analysis of Tweets in Three Indian Languages: Supplementary Material

1 Description

We include six tables here, corresponding to three languages (Bengali, Hindi, and Tamil), and 2-class and 3-class classification on them. All numbers reported are percentage accuracy values (micro-averaged) in 10-fold cross-validation on the training data of the corresponding language.

1. Table 1: Bengali, 2-class classification.
2. Table 2: Bengali, 3-class classification.
3. Table 3: Hindi, 2-class classification.
4. Table 4: Hindi, 3-class classification.
5. Table 5: Tamil, 2-class classification.
6. Table 6: Tamil, 3-class classification.

Feature Representation	Feature Category	Feature Type	NB	LR	DT	RF	SV	LS
	Surface features		51.66	53.88	48.81	50.87	52.61	49.29
Binary (Presence/Absence)	SentiWordNet		56.1	56.1	56.1	56.1	56.1	56.1
	Word unigrams	AW	67.67	64.03	59.75	59.75	56.1	64.03
		NS	65.61	62.28	56.89	60.22	56.1	62.28
		OS	56.26	56.89	55.94	56.74	56.1	56.89
	Word bigrams	AW	60.86	58.64	55.47	59.27	56.1	57.53
		NS	57.21	58.95	51.51	58.64	56.1	57.69
		OS	56.42	55.78	55.78	55.47	56.1	54.83
	Word trigrams	AW	57.37	59.43	51.19	53.25	56.1	58.95
		NS	56.26	59.43	53.09	53.09	56.1	58.95
		OS	56.1	56.1	55.94	56.1	56.1	56.1
	Character unigrams	AC	55.78	57.05	50.55	56.89	55.63	57.69
		SS	55.78	56.89	49.92	56.58	55.47	57.37
		PP	53.88	56.42	49.6	56.42	56.74	56.58
		SP	53.88	56.58	52.3	54.83	56.74	56.58
	Character bigrams	AC	51.82	53.09	53.88	58.16	56.1	51.66
		SS	51.19	52.93	55.78	56.89	56.1	52.3
		PP	52.46	53.41	54.04	58.8	56.1	54.2
		SP	52.14	54.52	58.8	59.75	56.1	54.68
	Character trigrams	AC	54.04	56.1	53.57	58.16	56.1	53.72
		SS	50.55	52.14	49.45	55.47	56.1	51.98
		PP	52.77	55.47	53.72	60.38	56.1	52.61
		SP	52.3	51.51	54.68	58.32	56.1	50.71
Tf (Term Frequency)	SentiWordNet		56.1	56.1	56.1	56.1	56.1	56.1
	Word unigrams	AW	67.35	64.66	59.43	59.59	56.1	64.82
		NS	66.88	62.76	58.48	59.9	56.1	62.12
		OS	57.05	58.16	55.31	56.58	56.1	57.37
	Word bigrams	AW	60.86	58.64	52.61	59.43	56.1	57.53
		NS	57.05	58.64	52.93	58.64	56.1	57.84
		OS	56.26	55.78	55.63	55.31	56.1	54.83
	Word trigrams	AW	57.53	59.43	50.71	53.25	56.1	58.95
		NS	56.26	59.43	50.87	53.09	56.1	58.95
		OS	56.1	56.1	56.26	56.1	56.1	56.1
	Character unigrams	AC	55.63	58.48	50.55	58.0	56.42	54.04
		SS	55.78	57.53	49.13	58.0	54.99	54.83
		PP	55.78	58.32	47.7	55.31	57.84	56.26
		SP	55.47	58.8	50.87	56.74	55.94	51.66
	Character bigrams	AC	53.57	54.83	51.66	57.21	56.1	53.41
		SS	51.51	54.68	51.35	57.21	55.94	53.41
		PP	53.09	53.72	53.57	57.69	55.94	52.93
		SP	53.09	55.31	55.63	58.8	56.1	52.77
	Character trigrams	AC	55.78	54.68	52.77	57.05	56.1	55.47
		SS	52.14	53.25	48.81	58.48	56.1	53.09
		PP	54.99	56.42	55.63	59.9	56.1	54.83
		SP	54.52	55.78	55.15	58.8	56.1	53.25
Tfidf (Term Frequency	SentiWordNet		56.1	56.1	56.1	56.1	56.1	56.1
Inverse Document Frequency)	Word unigrams	AW	67.67	64.18	60.38	59.59	56.1	63.39
		NS	67.83	64.34	59.11	59.9	56.1	63.39
		OS	57.05	57.53	54.36	56.58	56.74	57.21
	Word bigrams	AW	58.0	58.16	53.88	59.43	56.1	58.32
		NS	56.58	57.21	51.98	58.64	56.1	57.21
		OS	56.89	54.52	55.15	55.31	56.1	54.2
	Word trigrams	AW	57.21	58.95	54.04	53.25	56.1	54.04
		NS	56.26	58.95	50.55	53.09	56.1	53.72
		OS	57.69	56.1	56.1	56.1	56.1	56.1
	Character unigrams	AC	54.83	57.21	47.7	58.16	57.05	58.8
		SS	54.83	57.21	48.18	58.0	57.05	58.95
		PP	54.04	58.32	48.81	57.37	56.26	58.48
		SP	54.04	58.32	51.82	56.74	56.26	58.48
	Character bigrams	AC	51.66	56.89	49.13	57.21	55.31	57.37
		SS	51.19	56.74	51.35	57.21	55.47	56.1
		PP	48.65	50.55	53.09	57.69	55.94	49.92
		SP	51.35	55.15	55.47	58.8	55.94	52.77
	Character trigrams	AC	54.04	57.37	55.63	57.05	55.78	57.53
		SS	53.09	56.58	51.19	58.48	55.78	53.72
		PP	50.87	56.58	56.74	59.9	56.1	53.57
		SP	51.82	53.72	52.93	58.8	55.94	52.3

Table 1: % accuracy of 2-class classification for Bengali on 10-fold cross-validation on the training data. AW = all words, NS = all words except stop words, OS = only stop words. AC = all characters, SS = all characters except space, PP = all characters except punctuation, SP = all characters except space and punctuation.

Feature Representation	Feature Category	Feature Type	NB	LR	DT	RF	SV	LS
	Surface features		34.43	42.64	34.93	38.14	39.44	36.44
Binary (Presence/Absence)	SentiWordNet		36.84	36.84	36.84	36.84	36.84	36.84
	Word unigrams	AW	47.65	51.25	47.75	48.05	36.84	50.05
		NS	46.95	49.85	44.14	49.35	36.84	50.15
		OS	37.74	39.54	39.84	39.84	37.74	39.24
	Word bigrams	AW	45.45	48.55	44.74	46.05	36.84	48.15
		NS	43.84	47.45	44.04	45.75	36.84	47.55
		OS	36.14	36.04	37.64	38.14	36.84	35.44
	Word trigrams	AW	42.34	46.35	43.84	44.54	36.84	46.35
		NS	42.04	45.95	42.74	44.14	36.84	45.35
		OS	35.44	36.84	36.04	36.04	36.84	36.24
	Character unigrams	AC	39.14	43.34	38.14	45.95	40.24	43.54
		SS	39.24	43.44	38.04	44.84	40.24	43.34
		PP	36.74	41.34	39.54	43.14	40.04	41.54
		SP	36.84	41.14	39.34	43.34	40.04	41.54
	Character bigrams	AC	37.24	39.24	40.04	43.44	36.84	40.54
		SS	38.04	38.74	37.14	44.14	36.84	38.44
		PP	37.24	39.44	39.34	44.14	36.84	39.94
		SP	36.54	39.84	38.84	44.34	36.84	37.44
	Character trigrams	AC	38.24	41.04	39.14	44.44	36.84	40.94
		SS	36.94	39.44	35.74	45.25	36.84	37.04
		PP	36.84	38.34	37.34	47.75	36.84	37.84
		SP	37.64	38.04	41.54	44.94	36.84	36.14
Tf (Term Frequency)	SentiWordNet		36.84	36.84	36.84	36.84	36.84	36.84
	Word unigrams	AW	49.85	50.45	47.55	48.75	36.84	50.55
		NS	47.55	50.85	44.14	49.05	36.84	50.75
		OS	38.04	39.64	37.54	38.54	37.74	39.44
	Word bigrams	AW	45.45	48.75	46.15	46.15	36.84	48.45
		NS	43.84	47.55	44.54	44.54	36.84	47.65
		OS	36.44	35.64	37.84	38.04	36.84	35.54
	Word trigrams	AW	42.74	46.45	44.64	44.74	36.84	46.35
		NS	42.04	45.95	42.04	44.14	36.84	45.35
		OS	35.44	36.84	35.94	36.04	36.84	36.24
	Character unigrams	AC	41.54	45.95	38.84	44.94	43.14	39.84
		SS	41.84	46.05	39.14	44.84	43.44	38.44
		PP	40.14	44.04	39.94	44.04	43.54	41.14
		SP	39.84	44.84	39.54	43.44	44.24	40.54
	Character bigrams	AC	39.84	40.04	41.04	42.74	42.74	39.34
		SS	37.54	41.24	40.44	42.84	43.44	41.24
		PP	39.04	39.74	37.34	44.34	42.44	38.74
		SP	38.74	42.14	39.34	44.54	42.94	40.34
	Character trigrams	AC	40.14	40.84	38.34	43.94	38.54	40.84
		SS	38.24	42.64	39.64	43.74	38.94	41.14
		PP	40.14	41.54	37.34	46.75	39.24	40.94
		SP	40.54	43.14	38.54	46.15	39.14	40.84
Tfidf (Term Frequency Inverse Document Frequency)	SentiWordNet		36.84	36.84	36.84	36.84	36.84	36.84
	Word unigrams	AW	47.65	48.55	47.45	48.75	36.84	49.55
		NS	46.75	50.55	47.55	49.05	36.84	50.75
		OS	36.84	39.74	38.14	38.54	37.54	39.84
	Word bigrams	AW	41.64	47.75	44.74	46.15	36.84	48.15
		NS	39.24	47.85	43.84	44.54	36.84	45.65
		OS	36.34	36.24	37.84	38.04	36.84	36.34
	Word trigrams	AW	38.74	45.95	43.14	44.74	36.84	44.64
		NS	38.04	45.15	43.54	44.14	36.84	41.94
		OS	35.74	36.14	36.04	36.04	36.84	36.24
	Character unigrams	AC	39.24	45.45	40.94	45.15	43.84	44.24
		SS	39.24	45.45	41.74	44.84	43.74	44.54
		PP	37.54	42.54	40.44	44.04	39.94	41.34
		SP	37.54	42.54	40.84	43.44	40.04	41.54
	Character bigrams	AC	36.24	40.04	39.24	42.74	38.54	39.54
		SS	36.14	38.84	39.04	42.84	37.64	36.94
		PP	34.53	38.84	39.14	44.34	37.54	37.84
		SP	35.24	40.44	39.94	44.54	36.54	40.84
	Character trigrams	AC	38.24	41.64	39.74	43.94	36.64	40.64
		SS	37.84	41.24	40.34	43.74	36.54	41.14
		PP	36.54	41.24	36.94	46.75	36.94	38.74
		SP	36.64	41.74	38.04	46.15	36.64	39.94

Table 2: % accuracy of 3-class classification for Bengali on 10-fold cross-validation on the training data. AW = all words, NS = all words except stop words, OS = only stop words. AC = all characters, SS = all characters except space, PP = all characters except punctuation, SP = all characters except space and punctuation.

Feature Representation	Feature Category	Feature Type	NB	LR	DT	RF	SV	LS
	Surface features		64.79	75.79	66.99	75.1	78.4	65.75
Binary (Presence/Absence)	SentiWordNet		76.89	76.89	76.89	76.89	76.89	76.89
	Word unigrams	AW	78.68	81.57	76.62	79.92	76.89	80.61
		NS	73.04	80.33	75.93	79.78	76.89	79.78
		OS	73.59	74.97	67.68	76.89	76.89	71.53
	Word bigrams	AW	35.49	78.82	77.17	78.82	76.89	79.23
		NS	29.57	78.95	78.68	78.95	76.89	78.82
		OS	67.26	78.4	58.46	66.99	76.89	68.5
	Word trigrams	AW	28.2	78.68	78.54	78.82	76.89	78.82
		NS	30.4	78.68	78.82	78.82	76.89	78.95
		OS	46.22	78.4	77.44	75.93	76.89	76.48
	Character unigrams	AC	69.74	75.52	68.09	78.27	76.89	74.42
		SS	69.88	75.52	68.5	77.99	76.89	74.42
		PP	69.46	75.93	68.5	78.13	76.89	75.38
		SP	69.46	75.93	67.4	77.99	76.89	75.38
	Character bigrams	AC	75.24	76.48	71.11	78.82	76.89	70.29
		SS	75.1	75.24	69.19	78.13	76.89	71.11
		PP	74.42	77.17	70.84	78.95	76.89	73.18
		SP	74.83	75.93	69.88	78.4	76.89	72.21
	Character trigrams	AC	76.2	74.83	73.45	79.23	76.89	71.25
		SS	76.07	75.38	74.55	79.23	76.89	70.7
		PP	75.93	75.24	71.66	79.37	76.89	70.84
		SP	76.34	75.79	74.0	78.95	76.89	68.5
Tf (Term Frequency)	SentiWordNet		76.89	76.89	76.89	76.89	76.89	76.89
	Word unigrams	AW	78.4	79.78	75.24	79.37	76.89	77.99
		NS	71.66	80.06	76.75	79.5	76.89	79.78
		OS	71.94	75.52	68.64	77.99	76.89	71.94
	Word bigrams	AW	35.63	78.82	77.99	78.95	76.89	79.23
		NS	29.71	78.95	78.4	78.95	76.89	78.82
		OS	66.85	78.4	60.52	66.16	76.89	68.78
	Word trigrams	AW	28.2	78.68	78.82	78.82	76.89	78.82
		NS	30.4	78.68	78.68	78.82	76.89	78.95
		OS	46.22	78.4	77.03	76.07	76.89	76.48
	Character unigrams	AC	66.99	75.1	65.47	78.27	77.85	68.64
		SS	66.71	75.65	67.26	78.4	77.3	72.9
		PP	69.33	76.75	67.4	78.54	77.99	69.33
		SP	69.19	76.89	66.57	78.54	77.85	67.95
	Character bigrams	AC	73.04	72.9	68.78	78.68	76.89	69.88
		SS	73.04	74.97	67.26	77.85	76.89	71.8
		PP	72.9	74.55	69.05	78.4	76.89	71.8
		SP	73.59	77.58	68.36	78.13	76.89	69.88
	Character trigrams	AC	75.38	71.8	70.7	78.95	76.89	70.15
		SS	75.52	74.28	68.78	78.82	76.89	70.15
		PP	76.34	72.9	67.81	78.82	76.89	69.46
		SP	75.1	76.75	70.43	78.4	76.89	69.33
Tfidf (Term Frequency	SentiWordNet		76.89	76.89	76.89	76.89	76.89	76.89
Inverse Document Frequency)	Word unigrams	AW	53.23	81.43	76.34	79.37	76.89	79.64
		NS	48.56	80.74	76.75	79.5	76.89	80.33
		OS	62.04	69.74	67.68	77.99	77.03	69.74
	Word bigrams	AW	27.79	79.37	76.89	78.95	76.89	79.5
		NS	27.79	79.09	78.54	78.95	76.89	79.23
		OS	46.35	69.74	61.21	66.16	76.89	63.14
	Word trigrams	AW	27.79	79.09	77.99	78.82	76.89	78.82
		NS	27.1	78.82	78.54	78.82	76.89	78.82
		OS	31.64	75.24	77.3	76.07	76.89	76.07
	Character unigrams	AC	71.53	74.42	67.68	78.68	78.68	73.73
		SS	71.53	74.42	69.88	77.99	78.68	73.59
		PP	72.49	75.52	68.91	78.68	78.68	75.52
		SP	72.49	75.52	68.23	78.82	78.68	75.65
	Character bigrams	AC	65.61	71.94	67.81	78.95	77.03	67.26
		SS	66.02	73.04	68.09	78.54	77.03	66.71
		PP	65.47	72.76	67.26	78.82	77.03	69.6
		SP	65.34	72.76	67.81	78.4	77.03	68.91
	Character trigrams	AC	62.59	74.14	69.88	78.95	76.89	70.7
		SS	60.11	73.59	67.4	78.82	76.89	69.46
		PP	61.07	72.49	68.78	78.82	76.89	69.19
		SP	59.97	72.21	71.39	78.4	76.89	68.5

Table 3: % accuracy of 2-class classification for Hindi on 10-fold cross-validation on the training data. AW = all words, NS = all words except stop words, OS = only stop words. AC = all characters, SS = all characters except space, PP = all characters except punctuation, SP = all characters except space and punctuation.

Feature Representation	Feature Category	Feature Type	NB	LR	DT	RF	SV	LS
	Surface features		41.82	48.2	43.86	45.99	46.4	33.47
Binary (Presence/Absence)	SentiWordNet		45.74	45.74	45.74	45.74	45.74	45.74
	Word unigrams	AW	54.83	55.32	49.92	56.38	45.74	56.38
		NS	51.55	54.75	48.85	52.13	45.74	55.48
		OS	46.24	50.25	42.06	47.87	45.66	49.51
	Word bigrams	AW	25.2	52.45	48.28	47.95	45.74	52.86
		NS	21.19	46.07	43.45	47.71	45.74	46.64
		OS	44.68	50.49	46.15	48.12	45.74	47.3
	Word trigrams	AW	19.23	47.05	45.5	45.74	45.74	45.91
		NS	22.75	47.71	43.7	44.27	45.74	44.35
		OS	33.72	47.22	43.7	45.42	45.74	46.15
	Character unigrams	AC	45.25	50.98	42.39	51.31	49.59	50.82
		SS	45.17	50.98	43.13	52.45	49.59	50.82
		PP	45.25	49.26	43.37	50.41	49.75	48.61
		SP	45.01	49.18	41.0	51.47	49.84	48.61
	Character bigrams	AC	46.24	50.08	43.21	52.54	45.74	45.99
		SS	47.71	49.51	43.45	51.72	45.74	46.56
		PP	45.99	50.9	45.66	53.19	45.74	46.15
		SP	46.07	51.39	42.55	52.05	45.74	49.26
	Character trigrams	AC	48.04	47.38	45.17	53.03	45.74	45.5
		SS	47.38	47.05	43.86	52.29	45.74	46.64
		PP	48.2	48.2	42.31	53.11	45.74	45.42
		SP	46.89	49.02	44.84	51.8	45.74	46.15
Tf (Term Frequency)	SentiWordNet		45.74	45.74	45.74	45.74	45.74	45.74
	Word unigrams	AW	54.91	56.96	50.08	55.81	45.74	55.89
		NS	50.82	54.42	47.3	52.13	45.74	55.16
		OS	45.58	48.61	42.8	47.63	48.2	48.94
	Word bigrams	AW	25.2	52.54	47.55	49.35	45.74	53.11
		NS	21.19	46.32	43.29	47.71	45.74	46.64
		OS	43.62	49.02	46.32	49.35	45.74	46.24
	Word trigrams	AW	19.23	46.64	45.34	45.91	45.74	45.99
		NS	22.67	47.63	44.11	44.35	45.74	44.44
		OS	33.72	46.24	43.54	45.5	45.74	45.91
	Character unigrams	AC	43.78	49.18	41.9	50.57	48.04	44.44
		SS	43.45	49.59	43.62	49.59	47.63	45.17
		PP	45.58	48.28	42.72	50.16	47.55	41.98
		SP	45.66	48.04	43.04	50.65	47.38	43.78
	Character bigrams	AC	46.48	49.35	42.23	53.03	49.02	46.32
		SS	46.97	48.36	43.94	52.37	49.26	45.5
		PP	46.64	50.74	44.84	53.68	49.1	47.22
		SP	47.14	50.33	44.68	51.8	47.95	46.48
	Character trigrams	AC	49.35	50.49	43.86	52.7	49.18	47.95
		SS	47.79	48.77	43.94	51.72	49.18	44.76
		PP	50.33	49.67	44.84	53.27	49.26	46.97
		SP	48.12	49.51	43.7	51.47	48.85	46.32
Tfidf (Term Frequency	SentiWordNet		45.74	45.74	45.74	45.74	45.74	45.74
Inverse Document Frequency)	Word unigrams	AW	38.05	56.38	49.18	55.81	45.74	55.07
		NS	35.6	55.07	48.2	52.13	45.74	54.34
		OS	43.37	48.69	41.33	47.63	49.02	49.1
	Word bigrams	AW	17.27	53.68	47.22	49.35	45.74	53.19
		NS	17.02	47.14	43.62	47.71	45.74	46.64
		OS	31.26	46.07	45.99	49.35	45.74	45.01
	Word trigrams	AW	16.86	46.32	45.58	45.91	45.74	44.44
		NS	16.45	44.27	44.03	44.35	45.74	43.86
		OS	19.8	46.07	44.44	45.5	45.74	45.17
	Character unigrams	AC	45.17	50.08	43.37	50.9	50.08	49.43
		SS	45.17	50.08	44.68	51.39	50.0	49.51
		PP	46.24	48.69	44.68	51.15	51.8	48.94
		SP	46.15	48.69	44.44	51.31	51.8	49.35
	Character bigrams	AC	43.45	45.99	44.93	52.62	46.81	42.64
		SS	40.43	46.4	42.72	52.86	46.32	42.96
		PP	42.64	46.81	45.5	50.82	47.38	45.91
		SP	41.65	48.77	44.19	50.41	46.89	47.22
	Character trigrams	AC	41.0	46.81	43.7	52.7	46.15	43.86
		SS	39.69	46.97	43.21	51.72	46.15	44.03
		PP	40.26	48.45	43.94	53.27	46.07	45.34
		SP	38.95	46.89	42.96	51.47	46.15	44.68

Table 4: % accuracy of 3-class classification for Hindi on 10-fold cross-validation on the training data. AW = all words, NS = all words except stop words, OS = only stop words. AC = all characters, SS = all characters except space, PP = all characters except punctuation, SP = all characters except space and punctuation.

Feature Representation	Feature Category	Feature Type	NB	LR	DT	RF	SV	LS
	Surface features		53.06	56.47	50.64	54.48	54.34	51.92
Binary (Presence/Absence)	SentiWordNet		55.05	55.05	55.05	55.05	55.05	55.05
	Word unigrams	AW	62.16	60.17	56.76	57.61	55.05	59.46
	Word bigrams	AW	49.36	56.9	56.19	56.76	55.05	55.76
	Word trigrams	AW	44.95	57.04	56.76	56.9	55.05	57.18
	Character unigrams	AC	57.89	58.04	49.22	57.18	57.47	58.46
		SS	57.89	58.04	50.64	57.61	57.61	58.46
		PP	57.04	55.48	50.36	55.76	57.33	56.47
		SP	56.76	55.76	51.64	57.04	57.18	56.61
	Character bigrams	AC	58.32	57.18	53.49	59.89	55.05	52.35
		SS	59.46	57.04	49.93	59.17	55.05	53.63
		PP	55.48	54.62	52.49	59.74	55.05	54.91
		SP	55.76	55.33	53.91	59.89	55.05	55.62
	Character trigrams	AC	56.9	55.05	54.05	60.31	55.05	52.2
		SS	58.32	57.75	55.76	59.89	55.05	56.47
		PP	56.05	53.34	53.77	59.6	55.05	52.06
		SP	56.33	56.05	55.62	57.04	55.05	54.34
Tf (Term Frequency)	SentiWordNet		55.05	55.05	55.05	55.05	55.05	55.05
	Word unigrams	AW	61.74	60.17	58.89	59.03	55.05	58.61
	Word bigrams	AW	49.64	56.9	56.61	56.76	55.05	55.9
	Word trigrams	AW	44.81	57.04	56.61	56.9	55.05	57.18
	Character unigrams	AC	57.18	55.33	54.34	59.74	57.89	52.63
		SS	56.9	55.62	57.47	57.61	57.75	47.51
		PP	56.33	53.77	56.05	57.18	57.61	53.49
		SP	56.19	53.63	54.05	56.61	57.18	51.64
	Character bigrams	AC	59.03	56.19	55.9	59.46	58.46	52.35
		SS	56.47	57.75	53.49	58.89	58.61	53.49
		PP	57.61	55.05	50.36	59.17	58.61	52.63
		SP	56.76	56.05	52.35	58.46	58.04	54.91
	Character trigrams	AC	59.03	57.75	55.19	60.03	53.2	52.49
		SS	57.61	59.32	51.21	58.75	52.92	56.33
		PP	58.18	56.05	55.05	58.04	54.34	52.49
		SP	57.47	57.04	52.49	58.18	54.91	55.76
Tfidf (Term Frequency	SentiWordNet		55.05	55.05	55.05	55.05	55.05	55.05
Inverse Document Frequency)	Word unigrams	AW	58.32	59.6	56.61	59.03	55.05	58.75
	Word bigrams	AW	48.08	55.76	56.47	56.76	55.05	55.9
	Word trigrams	AW	46.09	57.04	56.61	56.9	55.05	55.76
	Character unigrams	AC	53.2	53.77	56.05	58.46	59.03	54.05
		SS	53.2	53.77	55.62	58.61	59.17	53.63
		PP	51.49	54.34	54.77	56.76	57.89	53.77
		SP	51.49	54.34	54.05	57.89	57.75	54.2
	Character bigrams	AC	57.04	54.2	53.06	59.03	54.77	50.64
		SS	56.61	55.62	53.91	59.03	54.91	53.63
		PP	55.33	54.91	50.21	57.75	54.77	57.18
		SP	54.34	56.33	55.76	58.61	54.77	56.76
	Character trigrams	AC	54.34	53.2	54.62	60.03	55.05	51.21
		SS	54.62	54.91	50.92	58.75	55.05	53.91
		PP	53.06	53.49	55.33	58.04	55.05	52.35
		SP	51.21	55.62	52.77	58.18	55.05	56.19

Table 5: % accuracy of 2-class classification for Tamil on 10-fold cross-validation on the training data. AW = all words (for Tamil, we did not have stop words). AC = all characters, SS = all characters except space, PP = all characters except punctuation, SP = all characters except space and punctuation.

Feature Representation	Feature Category	Feature Type	NB	LR	DT	RF	SV	LS
	Surface features		39.08	43.52	34.27	37.17	39.17	36.9
Binary (Presence/Absence)	SentiWordNet		36.26	36.26	36.26	36.26	36.26	36.26
	Word unigrams	AW	40.71	39.53	39.26	42.07	36.26	38.8
	Word bigrams	AW	36.08	40.25	38.08	39.26	36.26	40.34
	Word trigrams	AW	30.92	37.99	37.35	37.99	36.26	38.17
	Character unigrams	AC	43.52	40.98	36.08	43.16	43.25	40.71
		SS	43.52	41.07	36.9	44.15	43.34	40.34
		PP	43.25	41.25	39.98	41.07	43.16	40.62
		SP	43.43	41.25	39.44	39.89	43.06	40.34
	Character bigrams	AC	42.25	41.34	39.89	43.61	36.26	37.81
		SS	41.7	41.34	37.81	44.24	36.26	39.44
		PP	39.17	39.08	39.98	42.52	36.26	38.89
		SP	39.44	40.07	37.81	41.61	36.26	40.16
	Character trigrams	AC	39.26	38.8	37.53	43.34	36.26	35.63
		SS	39.89	41.98	37.53	43.16	36.26	39.89
		PP	38.44	39.26	37.99	40.89	36.26	38.53
		SP	38.62	40.98	35.9	41.98	36.26	38.89
Tf (Term Frequency)	SentiWordNet		36.26	36.26	36.26	36.26	36.26	36.26
	Word unigrams	AW	40.62	40.16	39.8	38.8	36.26	38.17
	Word bigrams	AW	36.08	40.44	39.08	39.17	36.26	39.89
	Word trigrams	AW	30.92	37.99	37.53	37.99	36.26	38.17
	Character unigrams	AC	44.24	40.07	37.08	45.24	39.71	37.99
		SS	44.24	40.16	35.99	43.79	39.44	36.9
		PP	43.79	39.62	36.36	42.43	38.44	37.26
		SP	43.97	39.71	34.81	44.15	39.62	36.45
	Character bigrams	AC	43.61	41.7	39.44	44.06	41.34	38.89
		SS	41.89	39.98	35.81	42.61	41.34	39.08
		PP	42.7	40.34	38.08	41.98	41.34	40.62
		SP	41.43	41.43	36.81	43.43	41.52	38.8
	Character trigrams	AC	40.62	40.62	36.81	40.34	40.89	37.35
		SS	40.16	42.43	36.9	42.52	40.89	40.25
		PP	39.98	41.43	35.99	39.62	40.34	38.8
		SP	40.16	41.98	38.8	40.34	39.98	41.89
Tfidf (Term Frequency	SentiWordNet		36.26	36.26	36.26	36.26	36.26	36.26
Inverse Document Frequency)	Word unigrams	AW	40.16	39.8	39.26	38.8	36.26	38.26
	Word bigrams	AW	33.18	41.25	38.62	39.17	36.26	40.89
	Word trigrams	AW	29.83	38.17	37.26	37.99	36.26	38.17
	Character unigrams	AC	40.98	40.53	36.63	43.79	41.8	40.8
		SS	40.98	40.53	36.36	44.51	41.8	40.98
		PP	40.44	40.25	34.81	41.8	43.34	40.34
		SP	40.44	40.25	35.27	43.16	43.16	39.98
	Character bigrams	AC	38.98	38.62	38.26	43.79	42.25	36.99
		SS	37.81	38.62	35.63	42.16	42.07	38.08
		PP	36.99	38.53	39.08	42.34	41.25	39.53
		SP	36.9	41.7	35.99	41.89	41.07	41.25
	Character trigrams	AC	37.53	38.35	36.63	40.34	40.34	35.54
		SS	37.08	39.35	38.53	42.52	38.53	36.63
		PP	36.54	39.35	36.54	39.62	40.25	37.9
		SP	35.9	41.34	39.44	40.34	38.26	38.89

Table 6: % accuracy of 3-class classification for Tamil on 10-fold cross-validation on the training data. AW = all words (for Tamil, we did not have stop words). AC = all characters, SS = all characters except space, PP = all characters except punctuation, SP = all characters except space and punctuation.